|  |
| --- |
| Trường Đại học Khoa học tự nhiên – Khoa Công nghệ thông tin. |
| Đồ án thực hành cuối kỳ |
| Nhập môn khoa học dữ liệu – Introduction to Data Science. |

|  |
| --- |
| Nhóm 9  Tháng 11, 2024. |

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**📖👨‍💻📖**

**A blue and white logo

Description automatically generated**

**ĐỒ ÁN THỰC HÀNH CUỐI KÌ**

**Bộ môn:** Nhập môn khoa học dữ liệu.

**Tên đề tài:**

*“Phân tích mối quan hệ giữa các loại gia vị, nguyên liệu và khả năng kết hợp của chúng trong các món ăn”.*

**STT nhóm:** 9.

**Thành viên:**

1. 22120384 – Nguyễn Đình Trí.
2. 22120398 – Vũ Hoàng Nhật Trường.
3. 22120412 – Nguyễn Anh Tường.
4. 22120424 – Nguyễn Ngọc Bảo Uyên.
5. 22120449 – Lê Nguyễn Huyền Vy.

# **Thông tin chung:**

1. **Bộ môn:** Nhập môn khoa học dữ liệu.
2. **Giảng viên lý thuyết:** Thầy Lê Ngọc Thành.
3. **Giảng viên thực hành:** Thầy Lê Nhựt Nam.
4. **Mã lớp:** 22\_21.
5. **STT nhóm:** 9.
6. **Danh sách thành viên:**
   1. 22120384 – Nguyễn Đình Trí.
   2. 22120398 – Vũ Hoàng Nhật Trường.
   3. 22120412 – Nguyễn Anh Tường.
   4. 22120424 – Nguyễn Ngọc Bảo Uyên.
   5. 22120449 – Lê Nguyễn Huyền Vy.
7. **Link github repository:** [“Click here to go to our github repository.”](https://github.com/richardnguyen0715/IntroDS-FinalPro-HandlingReal-WorldProblems)

MỤC LỤC

[**ĐỒ ÁN THỰC HÀNH SỐ 01** 2](#_Toc180870113)

[**Thông tin chung:** 3](#_Toc180870114)

[**Section 1:** *Introduction to the using Dataset.* 5](#_Toc180870115)

[**Giới thiệu về các thành phần được sử dụng.** 6](#_Toc180870116)

[**Yêu cầu 1:** 7](#_Toc180870117)

[**Yêu cầu 2:** 8](#_Toc180870118)

[**Yêu cầu 3:** 9](#_Toc180870119)

[**Yêu cầu 4:** 10](#_Toc180870120)

# **Section 0:***Bảng phân công công việc.*

1. **Bảng phân công công việc:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Công việc** | **Người thực hiện** | **Tiến độ** | **Đóng góp** | **Note** |
| Data Collection | Nguyễn Anh Tường | 100% | 90% | Các bạn chạy cào giúp vì bị trang web chặn. |
| Nguyễn Đình Trí | 2.5% |
| Hoàng Vũ Nhật Trường | 2.5% |
| Nguyễn Ngọc Bảo Uyên | 2.5% |
| Lê Nguyễn Huyền Vy | 2.5% |
|  | | | | |
| Data Preprocessing | Lê Nguyễn Huyền Vy | 90% | 95% |  |
| Nguyễn Anh Tường | 5% | Thảo luận cách làm |
|  | | | | |
| Data Exploration | Nguyễn Đình Trí | 0% |  | -- |
| Nguyễn Ngọc Bảo Uyên |  |
|  | | | | |
| Data Modeling & Evaluation | Hoàng Vũ Nhật Trường | 0% |  | -- |
| Nguyễn Anh Tường |  |

# **Section 01:** *Data Collection.*

***A blue line drawing of a server

Description automatically generated***

1. **Các trường dữ liệu cần cào:**
2. **Tên bài viết:**

**Ví dụ:**

* Title: Đổi Vị Với Phở Gà Trộn Lạ Miệng, Thơm Ngon Ăn Mãi Mà Không Thấy Ngán
* Title: Mì Ý Thịt Viên Sốt Cà Chua Nữ Hoàng Của Nền Ẩm Thực Nước Ý.
* Title: Khoai Tây Tẩm Gia Vị Nướng Kiểu Âu.

1. **Tên của các nguyên liệu của món ăn đó:**

**Ví dụ:**

Ingredients:

½ con gà ta

1 miếng gừng nhỏ (nạo vỏ, đập dập)

250ml nước

1 tsb muối

2 tbs xì dầu

2 tbs đường

1 tsp giấm tỏi

1 củ hành tây

500g bánh phở tươi

1 nắm giá

Rau mùi, hành lá

Hành phi

Lạc rang

Ingredients:

6-8 củ khoai tây

1 tbs dầu olive

½ tsp muối

½ tsp tiêu xay

½ bột ớt paprika

½ tsp bột tỏi

Hương thảo tươi

1. **Định dạng của dữ liệu sau khi cào về:**

Các món ăn hay bài viết khác nhau sẽ được phân tách với nhau bởi một dòng:

‘--------------------------------------------------‘

**Ảnh minh họa:**

**A screenshot of a computer program

Description automatically generated**

Figure . Ảnh minh họa cho tập dữ liệu sau khi cào.

1. **Quá trình thực hiện:**
2. **Kokotaru.com:**
3. **Nhận xét:** Sử dụng tương tác cuộn chuột để load trang và load bài viết.
4. **Chi tiết các bước thực hiện cào dữ liệu từ Kokotaru:**

Bước 1: Sử dụng selenium để giả lập trình duyệt chrome.

Bước 2: Sử dụng trình duyệt chrome với thao tác cuộn chuột qua tất cả các trang con của kokotaru.

Bước 3: Lấy hết tất cả các liên kết bài viết từ trang web.

A screenshot of a menu

Description automatically generated

Figure . Ảnh trang web kokotaru.

Bước 4: Truy cập vào từng bài viết và lấy ra thành phần của các món ăn.

1. **Một số vấn đề xảy ra:**

* Người viết bài và code web viết ở quá nhiều dạng trình bày khác nhau.
* Các bài viết không có nguyên liệu.
* Các bài viết nguyên liệu được chèn sẵn vào trong tiêu đề.
* Các bài viết nguyên liệu được trộn vào trong dòng văn giới thiệu.
* Các bài viết sử dụng nguyên liệu A,B,C nhưng không dùng đến.

1. **Kitchenart.com**
2. **Nhận xét:** Các trang được phân chia theo dạng bình thường, gồm 38 trang. Trong mỗi trang sẽ có đường dẫn đến trang tiếp theo.
3. **Chi tiết các bước cào dữ liệu:**

Bước 1: Thực hiện lấy hết đường dẫn dẫn đến các trang chứa bài viết.

Bước 2: Thực hiện vào từng trang tổng hợp lấy ra hết các đường dẫn dẫn đến các bài viết có trong trang đó.

Bước 3: Thực hiện vào từng đường dẫn của bài viết để lấy dữ liệu.

A screenshot of a computer

Description automatically generated

Figure . Ảnh mục nguyên liệu của kitchenart.com

A screenshot of a computer program

Description automatically generated

Figure . Mã HTML của phần nguyên liệu.

1. **Một số vấn đề xảy ra:**

* Trang web cấm truy cập với tốc độ rất nhanh và nhiều lần.
* Trung bình truy cập lần lượt 5-10 bài với tốc độ nhanh thì trang web đã phát hiện và tiến hành chặn truy cập. Đã thử truy cập 1 – 2 – 3 bài nhưng vẫn bị chặn.
* Thời gian chờ mở chặn rất lâu. Vẫn chưa có con số chính xác nhưng dao động trong khoảng từ 5 – 8 phút.

1. **Kết quả:**

**Kokotaru:** 399 / 468 bài viết.

**Kitchenart:** 682 / 760 bài viết.

**Tổng cộng:** 1081 bài viết.

# **Section 2:** *Data PreProcessing.*

***A blue line drawing of a server

Description automatically generated***